



# From Scarcity to Sufficiency: Speech Recognition Pipeline for Zero-resource Language

Nikolay Karpov<sup>1</sup>, Sofia Kostandian<sup>1</sup>, Nune Tadevosyan<sup>1</sup>, Alexan Ayrapetyan<sup>1</sup>, Andrei Andrusenko<sup>1</sup>,  
Ara Yeroyan<sup>2</sup>, Mher Yerznkanyan<sup>3</sup>, Vitaly Lavrukhin<sup>1</sup>

<sup>1</sup>NVIDIA, Armenia and USA

<sup>2</sup>Plat.ai, Armenia

<sup>3</sup>Buymie, Armenia

nkarpov@nvidia.com

## Abstract

The quality of Automatic Speech Recognition (ASR) systems largely depends on the availability of training data, which is predominantly accessible for either high-resource or low-resource languages. In contrast, languages such as Armenian face significant challenges due to the almost zero availability of public speech and text corpora. In this paper, we introduce a comprehensive framework that elevates data availability for a zero-resource language to a new level, thereby enabling the development of a fully operational online ASR model. Our approach involves data collection and processing through diverse resources, including audiobooks, paid crowdsourcing, and leveraging the volunteer platform to assemble a labeled dataset totaling 149 hours. This data made it possible to apply pseudo-labeling techniques on additional 145 hours of public audio data, achieving a new state-of-the-art Word Error Rate (WER) of 9.90% on Common Voice test. All datasets and ASR models are open-sourced.

**Index Terms:** Automatic Speech Recognition, Low-resource languages, data collection, Armenian language

## 1. Introduction

Automatic speech recognition (ASR) systems have become ubiquitous in everyday life. However, their stable and accurate operation requires large amounts of training data (labeled speech) that are predominantly available for high-resource languages such as Chinese, Spanish, English, Hindi, etc. In contrast, many low-resource languages suffer from a scarcity of high-quality speech data in the public domain, severely limiting the development of reliable ASR systems. For instance, the multilingual speech collection MOSEL includes only 17 hours of Irish and 20 hours of Danish labeled speech [1].

One approach to increasing training datasets is to incorporate unlabeled speech. In [2], the authors collected podcasts for the Greek language and subsequently generated pseudo-labels using WhisperX pipeline [3]. Similarly, other studies for low-resource languages have demonstrated the effectiveness of leveraging existing ASR models to produce pseudo-labels, artificial data augmentation, and the use of text-to-speech (TTS) systems [4–7]. [8] showed that low-resource languages can be integrated into pretrained multilingual ASR for unbalanced datasets using language-weighted dynamic cross-entropy and data augmentation. In [9] authors focused on adding new languages into pretrained Whisper for the ASR task.

This challenge is further exacerbated in zero-data scenario, where available public data is insufficient to train an initial ASR model. The Armenian language exemplifies this issue: at the time of starting this work, only 10 hours of open data were available comprising 3 hours from Mozilla Common Voice

(MCV) [10] and 7 hours from FLEURS [11].

The scarcity of public speech corpora for zero-resource languages (including Armenian) is primarily due to the challenges and high costs of recruiting native speakers, compounded by the limited availability of modern public text corpora for contributions on platforms such as MCV. Moreover, the popular Whisper model performs poorly for such languages, rendering it unsuitable for initial model training. In such scenarios, it becomes imperative to collect a sufficient quantity of high-quality transcribed speech data to establish a reliable baseline. For instance, the ASR model for Northern Sámi was initially trained using at least 27 hours of labeled speech, with further expansion by incorporating 99 hours of pseudo-labeled data [12].

Despite limited training data, several studies have already focused on developing ASR systems for the Armenian language [13–16]. Previous efforts to assemble an Armenian speech corpus include the ArmSpeech corpus [17, 18], which was compiled by sourcing high-quality audio samples from freely available audiobooks and real-life speech scenarios, yielding a total of 15.7 hours of data. However, the ArmSpeech corpus remains private and is not publicly accessible.

In this paper, we demonstrate the transition of data availability for a zero-resource language to a new level, enabling the development of a fully operational online ASR model for Armenian language. The proposed pipeline does not depend on the previous amount of data and could be applied even with zero starting data. We detail our process for data collection and processing from diverse sources, including audiobooks, MCV platform, and paid crowdsourcing initiatives. Our efforts yielded a dataset comprising 149 hours of labeled Armenian speech, which we used to train a stable ASR model. This initial model further enabled the generation of pseudo-labels from public audio data, ultimately allowing us to obtain new ASR models that significantly outperform existing baselines and achieve a new state-of-the-art word error rate of 9.9%. All labeled datasets and ASR models have been released into open source.

The primary contributions of this work are as follows:

- Labeled Armenian speech data (149 hours), along with text corpora under an open license and the associated data preparation pipelines.
- Open-sourced state-of-the-art ASR model for the Armenian language.<sup>1</sup>
- Comprehensive evaluation of the costs and benefits associated with different audio data sources for ASR training.

We believe that our work will be of significant value to the community, as it not only enhances resources for the Armenian language but also offers efficient tools and insights for scal-

<sup>1</sup>[https://huggingface.co/nvidia/stt\\_hybrid\\_fastconformer\\_hybrid\\_large\\_pc](https://huggingface.co/nvidia/stt_hybrid_fastconformer_hybrid_large_pc)

ing datasets in other low-resource languages. By systematically evaluating various data collection strategies and their impact on ASR performance, our approach offers a replicable framework for addressing data scarcity in speech recognition research.

## 2. Data Collection

The main difficulty in creating speech recognition technology for zero-resource languages usually lies in the small number of speakers of these languages. This leads to a scarcity of digital resources accessible online and a limited pool of individuals capable of labeling data, whether on a voluntary basis or for compensation.

The Armenian Wikipedia corpus had already been incorporated into the MCV text corpus, and no additional texts or labeled audio datasets were available under a permissive license. Moreover, Armenian is not part of any major language family and possesses distinct acoustic features, adding further challenges to the development process.

### 2.1. Volunteer Crowdsourcing

When we started developing Armenian ASR, the Common Voice corpus had just 5 hours of Armenian data, with only 3 hours validated. To address this, we first increased the number of texts available there by taking text from old books with permissive licenses. Then we organized two data collection events at the two largest universities in Armenia. Participants were invited to read aloud and record texts from Common Voice, or to validate the labeled data submitted by other contributors. To encourage participation, we offered certificates and exclusive custom merchandise.

During the two events, an additional 48 hours of speech data were collected, of which 47 hours were validated by the SDP<sup>2</sup>. The organizational expenses amounted to approximately \$1,000, and it took about a month of a manager’s work to complete the preparations.

These offline events were costly but are likely the most sustainable way to grow the dataset, as they raised awareness of the Common Voice platform for future contributions.

### 2.2. Open Audiobooks

The use of audiobooks as a means to extend our dataset is inspired by their success in a major speech recognition corpora such as LibriSpeech [19], which originates from the Gutenberg Project [20]. However, there are inherent challenges associated with audiobooks as they feature lengthy audio chapters, in contrast to ASR preferred segmented speech (for instance from 1 to 20 seconds long audio).

We used open-source package `vac_aligner`<sup>3</sup> previously proposed for Armenian audio-text pairs [16] to align and segment the long-duration Grqaser’s<sup>4</sup> audiobooks into manageable segments suitable for ASR training. This processing involved segmenting long audio chunks at punctuation marks and merging very short segments while ensuring smooth transitions through window functions and silence detection, preventing abrupt "audio shifts" or unnatural pauses. These books mostly contain classical literature, including famous novels, fairy tales, and folk tales. This approach enabled us to create an additional 22-

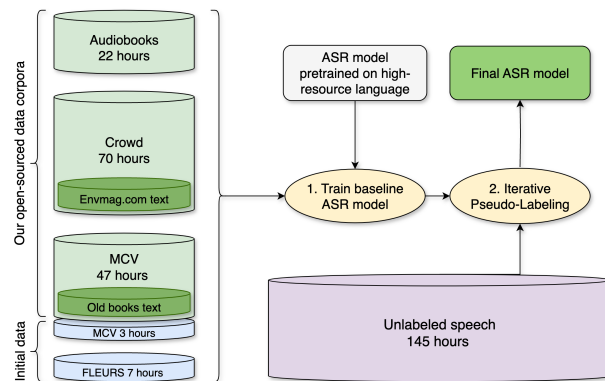


Figure 1: Illustration of the proposed ASR pipeline for data and model preparation.

hour corpus<sup>5</sup> for the Armenian speech, costing about one month of scientist’s work.

### 2.3. Paid Crowdsourcing

After leveraging volunteer crowdsourcing and audiobooks, we still had a lack of data. Therefore, we decided to use a crowdsourcing platform. For this, as described in previous work [21], we first need to obtain some texts and then vocalize them. During the volunteer crowdsourcing, we were relying on old books, which could be outdated. Instead, we have found a partner - Yerevan City Magazine, which agreed to share with the public their texts. We helped them to do it as part of the Yerevan Magazine Open Texts collection<sup>6</sup>.

Next, we processed these texts and used Toloka<sup>7</sup> platform for crowd-sourced reading. We implemented three stages: start, validate, and download using SDP pipelines<sup>8</sup>.

- **Start** by creating pool of workers who passed initial language test and submit them task to vocalize provided text.
- **Validate** the correctness of recorded utterances by ASR model and block dishonest workers.
- **Download** resulted audios with labels and convert them into needed format.

Reading correctness was validated automatically with our ASR model, trained on MCV and FLEURS datasets. The entire process took two months cost, around \$150, and required one month of engineering work. Ultimately, this cost-effective approach added 69.23 hours of verified speech data to our dataset.

### 2.4. Pseudo-labeling

To enhance the diversity and volume of speech data for ASR training, we identified suitable public Armenian audio content. We curated high-quality data, focusing on interviews, shows, and podcasts for their clarity and minimal noise, distributed under a permissive Creative Commons (CC BY) license.

The processing pipeline, implemented based on NeMo SDP<sup>9</sup>, begins with converting source audio files into single-channel 16,000 Hz WAV files. Once formatted, language identification [22] is performed to remove non-Armenian content

<sup>2</sup><https://github.com/NVIDIA/NeMo-speech-data-processor>

<sup>3</sup>VAC’s package - <https://pypi.org/project/vac-aligner/>

<sup>4</sup>A digital library of Armenian audiobooks - <https://grqaser.org>

<sup>5</sup><https://www.openslr.org/154>

<sup>6</sup><https://www.openslr.org/153/>

<sup>7</sup><https://toloka.ai/>

<sup>8</sup><https://github.com/NVIDIA/NeMo-speech-data-processor/pull/64>

<sup>9</sup><https://github.com/NVIDIA/NeMo-speech-data-processor>

#	Training subset	Size, h	Costs		MCV test			FLEURS test		
			Man-hour	Extra, \$	WER	WER noPC	PER	WER	WER noPC	PER
1	FLEURS + MCV5	12	0	0	65.39	64.59	62.90	83.67	82.38	53.43
2	FLEURS + Crowd	77	1 month	150	23.49	19.87	44.98	18.31	12.84	35.27
3	FLEURS + Audiobooks	29	1 month	0	41.71	38.26	56.53	26.03	20.12	41.53
4	FLEURS + MCV50	57	1 month	1,000	13.41	10.59	15.48	19.26	13.14	15.64
5	<i>Pseudo-labeled</i>	145	1 month	0	-	-	-	-	-	-
6	All available data	294	4 months	1,150	9.90	7.09	15.68	12.32	6.72	15.95

Table 1: Comparison of the collected datasets in terms of costs (Man-hour and extra cash payments) and the obtained speech recognition accuracy (WER, WER noPC, and PER in %) using FastConformer ASR models (RNN-T head).

and voice activity detection (VAD) using the *Marblenet* [23] model to isolate speech segments. Then NeMo’s force aligner is used for word-level predictions of border timestamps with our ASR model, trained on labeled datasets. This allows us to segment into chunks up to 20-seconds with intact sentences and words. Thus, silent and music segments were deleted, retaining only relevant speech data for ASR pseudo-labeling. This process yielded a cleaned dataset of 145 hours useful for self-training, costing us about one month of an engineering effort. In the Table 1 and 2 we call this subset as *Pseudo-labeled*.

### 3. Training Setup

All models were trained with NeMo framework and parameters from the standard configuration file<sup>10</sup>. To enhance the training process and achieve optimal results, some settings (listed below) were adapted for the Armenian language.

Key training parameters are 1,024 global batch size, 40 warmup steps, an initial learning rate of 0.005-0.006 with a cosine annealing scheduler, and training across 8 GPUs for 200 epochs. The initial model, pre-trained on English, is FastConformer Hybrid Large with around 114M parameters, which combines both RNN-T and CTC decoders, where the RNN-T head typically outperforms CTC as it is more context-aware and better suited for handling real-world speech data. We selected FastConformer Hybrid since it is the lightweight version of the top-ranked *parakeet-rmnt-0.6b*<sup>11</sup> and our goal is to develop accurate online model. Well-known Whisper based models are significantly larger and are considered as offline. Using such large models for low-resource languages risks overfitting and inefficient resource use during training and inference.

Each labeled subset is utilized with capital letters and following Armenian and common punctuation marks: “ ‘ ’ : - ‘ ‘ . . . . « » , ( ) . This improves the perception of the text and simplifies the inference pipeline since this eliminates the need for a separate model to predict them.

To ensure reliable evaluation and avoid inflated performance metrics, we carefully preprocess all datasets by removing extra characters and any overlap between training and testing splits for MCV and FLEURS datasets. After training, we average the weights of the five best checkpoints.

The model quality we measure using three metrics: Word Error Rate (WER), WER without punctuation marks and capitalized letters (WER noPC), and Punctuation Error Rate (PER)

<sup>10</sup>[https://github.com/NVIDIA/NeMo/blob/main/examples/asr/conf/fastconformer/hybrid\\_transducer\\_ctc](https://github.com/NVIDIA/NeMo/blob/main/examples/asr/conf/fastconformer/hybrid_transducer_ctc)

<sup>11</sup>[https://huggingface.co/spaces/hf-audio/open\\_asr\\_leaderboard](https://huggingface.co/spaces/hf-audio/open_asr_leaderboard)

after RNN-T head. PER metric was recently proposed by Meister et al. [24] to focus only on a predetermined subset of tokens, specifically punctuation marks:

$$PER = \frac{I_P + D_P + S_P}{I_P + D_P + S_P + C_P} \quad (1)$$

where each operation is related to the group of punctuation tokens only.  $D_P$ ,  $I_P$ ,  $S_P$  - instances where a token from the target group is deleted, inserted, or substituted relative to the reference sequence.  $C_P$  - instances where tokens from the target group exist at identical positions in both the reference and hypothesis sequences.

For each pair of WER values in Table 1 and 2 we calculated a pairwise confidence interval and determined the probability of improvement (POI) as defined by M. Bisani [25]. All POI values exceeded the 95% threshold, except for the pair corresponding to row 3 versus row 5 in Table 2, where the POI is approximately 0.85, which is still acceptable.

## 4. Experiment Results

### 4.1. Volunteer Crowdsourcing

First, we developed models using datasets FLEURS and Common Voice with 50 hours - MCV50. A comparison with the previous version of MCV5 - about 5 hours, demonstrates drastic improvement from 65.39 and 83.67 to 13.41 on MCV and 19.26 on FLEURS test. Row MCV50 + FLEURS in Table 2 is considered a baseline to count relative improvements for further experiments.

### 4.2. Open Audiobooks

Audiobooks tend to be recordings of professional speakers in clean environments and speech with specific prosody, intonation, and style that differ from ordinary native speakers. So, training on that and evaluating using MCV and FLEURS is definitely a huge domain shift. Unsurprisingly, you can find low performance of corresponding model Audiobooks + FLEURS in Table 1. Combining Audiobooks with a small MCV50 set, as noted in [16], reduces the model’s performance on the MCV test set.

In our experiments, we used a bigger training corpus with two extra sources - FLEURS and Crowd. Adding Audiobooks helped to decrease the test WER values - Table 2 row MCV50 + FLEURS + Crowd versus row MCV50 + FLEURS + Crowd + Audiobooks. This demonstrates that audiobooks themselves are valid and relevant sources, just need an appropriate amount of mixed data to reach better generalization.

For a better illustration, we conducted one more experiment with vs without using Audiobooks. Looking into Table 2 row

#	Training subset	Size, h	MCV test			FLEURS test		
			WER	WER noPC	PER	WER	WER noPC	PER
1	FLEURS + MCV50	57	13.41	7.09	15.48	19.26	13.14	15.64
2	FLEURS + MCV50 + <i>Pseudo-labeled</i>	202	12.81	9.85	15.38	16.17	9.70	17.05
3	FLEURS + MCV50 + Crowd	127	10.83	8.04	15.78	14.56	8.80	16.01
4	FLEURS + MCV50 + Crowd + <i>Pseudo-labeled</i>	272	10.62	9.03	15.63	13.67	8.17	15.9
5	FLEURS + MCV50 + Crowd + Audiobooks	149	10.80	7.39	<b>14.96</b>	13.03	7.40	<b>15.54</b>
6	FLEURS + MCV50 + Crowd + Audiobooks + <i>Pseudo-labeled</i>	294	<b>9.90</b>	<b>7.09</b>	15.68	<b>12.32</b>	<b>6.72</b>	15.95
7	Whisper_Large_v3		54.17	50.84	23.32	45.60	40.75	33.62

Table 2: Cumulative data extension of the training set and its impact on WER, WER noPC, and PER values (in %) for FastConformer ASR models (RNN-T head)

MCV50 + FLEURS + Crowd + Audiobooks + *Pseudo-labeled*, where we added Audiobooks on top of the training dataset used in row MCV50 + FLEURS + Crowd + *Pseudo-labeled* - we observe a notable enhancement evidenced by a significant 21% decrease from 10.62 to 9.9 WER on the MCV and a 17% decrease from 13.67 to 12.32 WER on the FLEURS test sets.

Hence, the Audiobooks helped to improve the results on MCV test (dataset from other domain), yielding the best results among all the experiments, through a better model generalization. Therefore, we helped the authors from Grqaser to make the constructed dataset (21.96 hours of audio-text pairs) accessible to the general public via publishing in OpenSLR.

### 4.3. Paid Crowdsourcing

A model training results on the almost 70 hours of crowd-sourced data, which is the biggest part of our labeled set, can be seen in Table 1 as Crowd + FLEURS. It demonstrates the best FLEURS WER value among lines from 1 to 4 (single training set). We attribute this high performance not only to the substantial data size but also to the high quality of speech produced by modern text and the advanced data preparation pipeline.

Mixing this data with the Common Voice training set MCV50 + FLEURS + Crowd in Table 2 improved WER values for both the MCV test from 13.41 to 10.83, which is a 19% relative improvement, and the FLEURS test from 19.26 to 14.56 (24% relatively).

### 4.4. Pseudo-labeling

We applied three experiments to explore how iterative pseudo-labeling (IPL) [26] algorithm with unlabeled data could improve ASR performance. Each experiment utilizing the same amount of *Pseudo-labeled* data but gradually increasing the size of labeled set MCV50 + FLEURS from 57 hours to MCV50 + FLEURS + Crowd + Audiobooks set with 149 hours. Relative improvement on MCV have gradually increased with the size of data from 4% to 26% and on FLEURS from 16% to 36%. These results clearly demonstrate that the ratio of labeled and unlabeled data has significantly affected the WER improvement, emphasizing its effectiveness in leveraging unlabeled data for low-resource languages.

It is expected that the best WER scores was achieved on the largest labeled set along with the unlabeled portion. Moreover, the PER metric for setups with *Pseudo-labeled* data was consistently worse than for those without it on the FLEURS test set. This is due to the significant differences between punctuation of snatches of phrases from *Pseudo-labeled* and complete expressions from FLEURS.

### 4.5. Costs and Benefits Evaluation

Finally, let us evaluate the costs and benefits of various audio data sources, including public audiobooks, volunteer-based and paid crowdsourcing, as well as pseudo-labeling. We spent approximately the same amount of time on each data source - about one month of working a full-time specialist, but received various amounts of the audio from 22 to 70 hours. This led us to various quality of ASR ranging from 18.31 to 26.03 of FLEURS WER % (Table 1). For the crowdsourcing, we incurred additional costs of \$1,000 and \$150.

From our experience, the best price-quantity ratio can be achieved by using the paid crowdsourcing method (70 hours). Together with the following self-training by iterative pseudo-labeling technique (Table 2), it demonstrates the best WER score on the FLEURS test set. Surprisingly, volunteer crowdsourcing showed a higher cost for less data (47 hours). The audiobooks processing is easily scalable; however, in our case, it yielded a smaller amount of data (22 hours), primarily due to the limited availability of resources in the specific language. As anticipated, merging all data sources results in the best performance with respect to WER values.

By utilizing the proposed pipeline, the reader can select the most suitable data source for the target language and effectively expand the dataset to a size large enough to train an ASR model with high accuracy.

## 5. Conclusion

Our work addresses the scarcity of Armenian ASR datasets by developing a scalable pipeline that combines multiple data collection strategies — volunteer and paid crowdsourcing, audiobooks, and pseudo-labeling. This approach resulted in a high-quality 149-hour labeled dataset and a state-of-the-art ASR model with a 9.90% WER, significantly improving Armenian ASR performance. Beyond its immediate impact, our findings offer a replicable framework for other zero-resource languages by systematically evaluating the trade-offs between cost, scalability, and model accuracy.

We moved the Armenian language from zero-resource to low-resource category and evaluated it against a proprietary model, Whisper Large v3, in a zero-shot manner without any fine-tuning to establish baselines. This revealed the limitations of current methods for low-resource languages. Further steps could be to finetune these models using newly collected data and apply data augmentation techniques suitable for low-resource languages.

## 6. References

- [1] M. Gaido, S. Papi, L. Bentivogli, A. Brutti, M. Cettolo, R. Gretter, M. Matassoni, M. Nabih, and M. Negri, “Mosel: 950,000 hours of speech data for open-source speech foundation model training on eu languages,” *arXiv preprint arXiv:2410.01036*, 2024.
- [2] G. Paraskevopoulos, C. Tsoukala, A. Katsamanis, and V. Katsouros, “The greek podcast corpus: Competitive speech models for low-resourced languages with weakly supervised data,” in *Interspeech 2024*, 2024, pp. 3969–3973.
- [3] M. Bain, J. Huh, T. Han, and A. Zisserman, “Whisperx: Time-accurate speech transcription of long-form audio,” *INTER-SPEECH 2023*, 2023.
- [4] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, “Making more of little data: Improving low-resource automatic speech recognition using data augmentation,” *arXiv preprint arXiv:2305.10951*, 2023.
- [5] A. Laptev, A. Andrusenko, I. Podluzhny, A. Mitrofanov, I. Medennikov, and Y. N. Matveev, “Dynamic acoustic unit augmentation with bpe-dropout for low-resource end-to-end speech recognition,” *Sensors (Basel, Switzerland)*, 2021.
- [6] A. Ullah, A. Ragano, and A. Hines, “Reduce, reuse, recycle: Is perturbed data better than other language augmentation for low resource self-supervised speech models,” in *Interspeech 2024*, 2024, pp. 77–81.
- [7] K. S. Bhogale, D. Mehendale, N. Parasa, T. Javed, P. Kumar, M. M. Khapra *et al.*, “Empowering low-resource language asr via large-scale pseudo labeling,” *arXiv preprint arXiv:2408.14026*, 2024.
- [8] A. Piñeiro-Martín, C. García-Mateo, L. Docio-Fernandez, M. del Carmen López-Pérez, and G. Rehm, “Weighted cross-entropy for low-resource languages in multilingual speech recognition,” in *Interspeech 2024*, 2024, pp. 1235–1239.
- [9] M. Qian, S. Tang, R. Ma, K. M. Knill, and M. J. Gales, “Learn and don’t forget: Adding a new language to asr foundation models,” in *Interspeech 2024*, 2024, pp. 2544–2548.
- [10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.
- [11] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech,” *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2022.
- [12] Y. Getman, T. Grosz, K. Hiovain-Asikainen, and M. Kurimo, “Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sámi,” in *Interspeech 2024*, 2024, pp. 2539–2543.
- [13] A. Vardanyan, “Noise-robust speech recognition system for armenian language,” Ph.D. dissertation, Master’s thesis, American University of Armenia, 2016.
- [14] S. Chakmakjian and I. Wang, “Towards a unified asr system for the armenian standards,” in *DIGITAM*, 2022.
- [15] V. Baghdasaryan, “Armenian speech recognition system: Acoustic and language models,” *International Journal Of Scientific Advances*, 2022.
- [16] A. Yeroyan and N. Karpov, “Enabling asr for low-resource languages: A comprehensive dataset creation approach,” *arXiv preprint arXiv:2406.01446*, 2024.
- [17] V. H. Baghdasaryan, “Armspeech: Armenian spoken language corpus,” *International Journal of Scientific Advances*, vol. 3, no. 3, pp. 454–459, 2022.
- [18] V. Baghdasaryan, “Extended armspeech: Armenian spoken language corpus,” *International Journal Of Scientific Advances*, 2022.
- [19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [20] B. Stroube, “Literary freedom: Project gutenber,” *XRDS: Crossroads, The ACM Magazine for Students*, vol. 10, no. 1, pp. 3–3, 2003.
- [21] N. Karpov, A. Denisenko, and F. Minkin, “Golos: Russian Dataset for Speech Research,” in *Proc. Interspeech 2021*, 2021, pp. 1419–1423.
- [22] F. Jia, N. R. Koluguri, J. Balam, and B. Ginsburg, “Ambernet: A compact end-to-end model for spoken language identification,” *arXiv preprint arXiv:2210.15781*, 2022.
- [23] F. Jia, S. Majumdar, and B. Ginsburg, “Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6818–6822.
- [24] A. Meister, M. Novikov, N. Karpov, E. Bakhturina, V. Lavrukhin, and B. Ginsburg, “Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.
- [25] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in asr performance evaluation,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. 1–409.
- [26] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, “slimipl: Language-model-free iterative pseudo-labeling,” *arXiv preprint arXiv:2010.11524*, 2020.